PODSTAWY ANALIZY DANYCH POMIAROWYCH CZ. 2

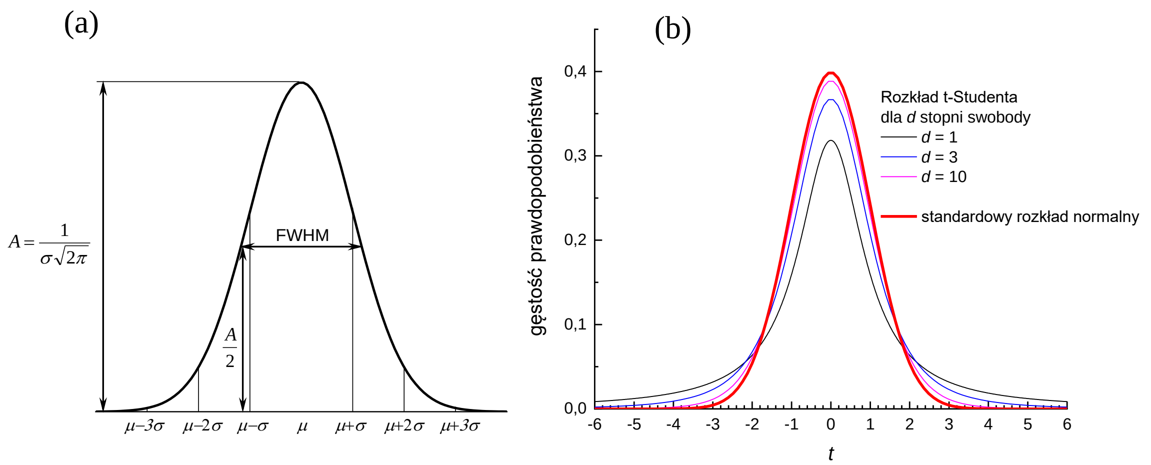
Rozkład normalny

Jednym z najważniejszych rozkładów prawdopodobieństwa jest rozkład normalny. Jest on bardzo często obserwowany, gdyż, jeśli jakaś wielkość jest sumą lub średnią bardzo dużej liczby zmiennych losowych, to niezależnie od rozkładu każdego z tych czynników jej rozkład będzie zbliżony do normalnego. W szczególności wielokrotne powtarzanie tego samego pomiaru daje wyniki rozrzucone wokół określonej wartości. Jeśli wyeliminujemy wszystkie większe przyczyny błędów, zakłada się, że pozostałe mniejsze błędy muszą być rezultatem dodawania się do siebie dużej liczby niezależnych czynników, co daje w efekcie rozkład normalny. Odchylenia od rozkładu normalnego rozumiane są jako wskazówka, że zostały pominięte błędy systematyczne. To stwierdzenie jest głównym założeniem teorii niepewności pomiarowych. Choć rozkład normalny jest często stosowanym założeniem, nigdy nie jest ściśle realizowany. Ma on bowiem niezerową gęstość prawdopodobieństwa dla dowolnej wartości zmiennej losowej, podczas gdy w rzeczywistości zmienne są zawsze ograniczone, a często nieujemne. Mimo to rzeczywisty rozkład jest często bardzo zbliżony do normalnego, stąd zwykle zakłada się, że zmienna ma rozkład normalny. Nie należy jednak robić tego bez sprawdzenia jak wielkie są rozbieżności. Rozkłady dalekie od normalnego (np. z elementami odstającymi) mogą sprawić, że wyniki metod statystycznych będą mylnie interpretowane. Przykładem są tu metody regresji liniowej oraz korelacji Pearsona, które, choć zdefiniowane dla dowolnych rozkładów, mają sensowną interpretację tylko dla wielowymiarowego rozkładu normalnego wektora próbki. Jeśli w próbce występują na przykład elementy odstające, co jest szczególnym przypadkiem rozkładu dalekiego od normalnego, korelacja może przyjąć dowolną wartość między -1 a $+1$, bez względu na rzeczywistą zależność między zmiennymi losowymi. Także regresja będzie dawała błędne rezultaty.

Funkcja gęstości prawdopodobieństwa rozkładu normalnego ze średnią μ i odchyleniem standardowym σ jest przykładem funkcji Gaussa opisanej wzorem:

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

Wykres funkcji prawdopodobieństwa tego rozkładu (Rysunek 1 a) jest krzywą w kształcie dzwonu, symetryczną względem wartości średniej rozkładu, której szerokość połówkowa (FWHM – *Full Width at Half Maximum*) wynosi $2\sigma\sqrt{\ln 4}$. Punkt przegięcia krzywej znajduje się w odległości jednego odchylenia standardowego od średniej.



Rysunek 1. (a) Wykres gęstości prawdopodobieństwa rozkładu normalnego oraz (b) porównanie standardowego rozkładu normalnego z rozkładem t-Studenta w zależności od liczby stopni swobody.

Około 68,3% pola pod wykresem krzywej znajduje się w odległości jednego odchylenia standardowego od średniej, około 95,5% w odległości dwóch odchyleń standardowych i około 99,7% w odległości trzech.

Jeśli $\mu = 0$ i $\sigma = 1$, to rozkład ten nazywa się standardowym rozkładem normalnym z funkcją gęstości prawdopodobieństwa opisywaną przez

$$f = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \quad (2)$$

Oznacza to, że jeśli x jest zmienną losową o rozkładzie normalnym ze średnią μ i odchyleniem standardowym σ, to $z = \frac{x-\mu}{\sigma}$ jest zmienną losową o standardowym rozkładzie normalnym.

Standardowy rozkład normalny został stablicowany i inne rozkłady normalne są prostymi transformacjami rozkładu standardowego. W ten sposób możemy używać tablic dystrybuanty standardowego rozkładu normalnego do wyznaczenia wartości dystrybuanty rozkładu normalnego o dowolnych parametrach.

Test t-Studenta

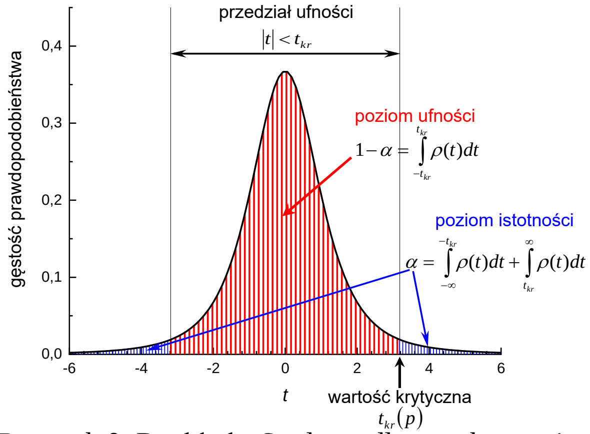
Ważnym elementem analizy danych doświadczalnych jest sprawdzanie zgodności wyników doświadczeń z przewidywaniami teoretycznymi lub też sprawdzanie wzajemnej zgodności wyników różnych pomiarów, a więc, mówiąc ogólnie, testowanie hipotez. Test t-Studenta jest jednym z mniej skomplikowanych i bardzo często wykorzystywanych testów statystycznych używanych do weryfikacji hipotez gdy mamy do czynienia z małymi próbami (najczęściej przyjmuje się, że próba jest mała gdy jej liczebność jest mniejsza niż 30, przy większych liczebnościach przyjmuje się, że zbiega on do rozkładu normalnego – Rysunek 1 b). Jest na przykład bardzo często wykorzystywany do porównania dwóch średnich między sobą. Ważne podkreślenia jest, że **test t-Studenta jest testem istotności, tzn. rodzajem testu, w którym na podstawie wyników próby losowej podejmuje się wyłącznie decyzję odrzucenia hipotezy, którą się sprawdza, bądź stwierdza się brak podstaw do odrzucenia tej hipotezy**. Nie podejmuje się natomiast decyzji o przyjęciu sprawdzanej hipotezy, ponieważ bierze się w tym teście pod uwagę tylko **błąd pierwszego rodzaju** (błąd polegający na odrzuceniu hipotezy, która w rzeczywistości nie jest fałszywa), nie uwzględnia się natomiast konsekwencji popełnienia błędu drugiego rodzaju (błędu polegającego na nieodrzuceniu hipotezy, która jest w rzeczywistości fałszywa).

1. **Test t-Studenta dla jednej próby** określa, czy średnia z próbki pobranej z populacji o normalnym rozkładzie jest zgodna z wartością hipotetyczną dla danego poziomu istotności. Jeśli zmienne losowe $X_1, X_2, \dots, X_n$ mają jednakowy rozkład prawdopodobieństwa, który jest rozkładem normalnym o średniej μ i odchyleniu standardowym σ, wówczas zmienna:
$$t = \frac{\hat{X}-\mu}{s_X\sqrt{\frac{1}{n}}} \quad (3)$$
ma rozkład t-Studenta o $d = n - 1$ stopni swobody. $\hat{X}$ jest wartością średnią z próby, natomiast s_X – odchyleniem standardowym z eksperymentalnym.
2. **Test t-Studenta dla dwóch prób niezależnych** określa, czy średnie z dwóch niezależnych próbek o rozkładzie normalnym są równe, czy też różnią się od siebie i określa poziom istotności różnicy średnich. Jeśli dwie próby o liczebności n_1 i n_2, wartości średnich $\hat{X}_1$ i $\hat{X}_2$ i odchyleniach standardowych eksperymentalnych s_1 i s_2 zostały wylosowane z populacji mających taki sam rozkład normalny, wówczas zmienna:
$$t = \frac{\hat{X}_1-\hat{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \quad (4)$$
ma rozkład t-Studenta o $d = n_1 + n_2 - 2$ stopni swobody, gdzie $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

Aby rozstrzygnąć czy wartość t-Studenta jest istotna statystycznie, to znaczy wskazuje na istotną statystycznie różnicę między $\hat{X}_1$ i $\hat{X}_2$ (czy $\hat{X}$ i μ), należy określić wymagany poziom istotności prowadzonego testu (Rysunek 2). **Poziom istotności** α wyraża prawdopodobieństwo popełnienia błędu I-go rodzaju (błędu polegającego na odrzuceniu hipotezy, która w rzeczywistości nie jest fałszywa) i określa maksymalne ryzyko błędu jakie jesteśmy skłonni zaakceptować przyjmując hipotezę. Z poziomem istotności α związane są następujące pojęcia:

poziom ufności zdefiniowany jako $1-\alpha$, oraz **przedział ufności**, taki przedział parametru t w którym na $(1-\alpha)\cdot 100\%$ leży rzeczywista wartość parametru t;

Wybór poziomu istotności zależy od badacza, natury problemu i od tego jak dokładnie chce on weryfikować swoje hipotezy. Najczęściej przyjmuje się poziom istotności 0,05, 0,03 lub 0,01. Hipotezy przyjmujemy jako niesprzeczną z danymi, ze współczynnikiem ufności $1-\alpha$, jeśli obliczona wartość t mieści się w przedziale ufności. Im większy poziom ufności, tym szerszy przedział ufności, a więc tym mniej godzimy się na popełnienie błędu pierwszego rodzaju.



Rysunek 2. Rozkład t-Studenta dla trzech stopni swobody z zaznaczoną wartością krytyczną, przedziałem ufności oraz współczynnikiem ufności, dla poziomu istotności $\alpha = 0{,}05$.

Wartości krytyczne, tzn. granice przedziałów ufności dla zadanej liczby stopni swobody (d) oraz poziomu istotności (α) są stabelaryzowane (Tabela 1). Aby sprawdzić, czy obliczona wartość t-Studenta wskazuje na istotną statystycznie zależność, musimy sprawdzić czy znajduje się w przedziale ufności, tzn. porównać wartość bezwzględną otrzymanej wartości t z wartościami krytycznymi rozkładu dla zadanego poziomu istotności. Jeżeli wartość bezwzględna obliczonej wartości testu t-Studenta w naszym badaniu jest większa od wartości krytycznej oznacza to, że istotność statystyczna testu jest mniejsza niż przyjęty poziom istotności co uprawnia do odrzucenia hipotezy (na zadanym poziomie istotności). Jeżeli zaś wartość bezwzględna obliczonej wartości testu t-Studenta w naszym badaniu jest mniejsza od wartości krytycznej nie mamy podstaw do odrzucenia hipotezy. Jeśli dysponujemy oprogramowaniem, który umożliwia wyznaczanie dystrybuanty rozkładu (całki rozkładu prawdopodobieństwa), lepiej jest, w miejsce wartości krytycznej, podawać tzw. prawdopodobieństwo testowe (wartość *p*, p-value) – prawdopodobieństwo kumulatywne wylosowania próby takiej lub bardziej skrajnej jak zaobserwowana. Wartość *p* testu odpowiada takiemu poziomowi istotności, przy którym dla zaobserwowanej wartość statystyki testowej zmienia się konkluzja testu. Jeśli wartość *p* testu jest niższa, niż przyjęty z góry poziom istotności statystycznej można postępować tak, jakby hipoteza została odrzucona.

Tabela 1. Wartości krytyczne rozkładu t-Studenta dla zadanej liczby stopni swobody d i wybranych poziomów istotności α.

α \ d	1	2	3	4	5	6	7	8	9	10
0,01	63,6567	9,9248	5,8409	4,6041	4,0321	3,7074	3,4995	3,3554	3,2498	3,1693
0,03	21,2050	5,6428	3,8961	3,2976	3,0029	2,8289	2,7146	2,6338	2,5738	2,5275
0,05	12,7062	4,3029	3,1824	2,7764	2,5706	2,4469	2,3646	2,306	2,2622	2,2281
α \ d	11	12	13	14	15	16	17	18	19	20
0,01	3,1058	3,0545	3,0123	2,9768	2,9467	2,9208	2,8982	2,8784	2,8609	2,8453
0,03	2,4907	2,4607	2,4359	2,4149	2,3970	2,3816	2,3681	2,3562	2,3457	2,3362
0,05	2,201	2,1788	2,1604	2,1448	2,1314	2,1199	2,1098	2,1009	2,093	2,086